UNIVERSITÄT AUGSBURG
Universitätsbibliothek

OPEN ACCESS
OPUS AUGSBURG

Towards vision-language models for cardiac surgery lecture videos

Georgii Kostiuchik, Jimmy Chen, Mina Farag, Gabriele Romano, Matthias Karck, Rawa Arif, Philippe Grieshaber, Sandy Engelhardt

Angaben zur Veröffentlichung / Publication details:

Kostiuchik, Georgii, Jimmy Chen, Mina Farag, Gabriele Romano, Matthias Karck, Rawa Arif, Philippe Grieshaber, and Sandy Engelhardt. 2025. "Towards vision-language models for cardiac surgery lecture videos." *Current Directions in Biomedical Engineering* 11 (2): 41–44. https://doi.org/10.1515/cdbme-2025-0311.

Nutzungsbedingungen / Terms of use:

CC BY 4.0

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
CC-BY 4.0: Creative Commons: Namensnennung
Weitere Informationen finden Sie unter: / For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Georgii Kostiuchik*, Jimmy Chen, Mina Farag, Gabriele Romano, Matthias Karck, Rawa Arif, Philippe Grieshaber, and Sandy Engelhardt

Towards vision-language models for cardiac surgery lecture videos

https://doi.org/10.1515/cdbme-2025-0311

Abstract: Vision-language models represent an emerging paradigm that leverages natural language to train vision systems with broad capabilities. Recently, the use of surgical lecture videos has emerged as a promising method for developing models capable of understanding surgical scenes. In this work, we aim to translate these developments to the domain of cardiac surgery, which is marked by heterogeneity and complexity of surgical cases. To this end, we curate a dataset of cardiac surgery lecture videos and augment the training dataset by using a Large Language Model (LLM) to extract procedural steps for each surgery. Preliminary results suggest that this form of data augmentation can enhance model performance on text-based video retrieval tasks.

Keywords: Vision-Language Models, Vision-Language Pretraining, Surgical Computer Vision

1 Introduction

Vision-language models offer promising new opportunities for the surgical domain by enabling the joint analysis of text and images, while supporting a wide range of downstream tasks. CLIP (Contrastive Language-Image Pretraining) [1] introduced a novel approach of leveraging natural language as supervision for training vision models. SurgVLP [2] represents one of the first efforts to train a vision-language model on laparoscopic lecture videos within the surgical domain. The authors demonstrate that the model can perform key surgical computer vision tasks (e.g., surgical phase recognition) without being explicitly trained for these tasks. Since then, several follow-up works have been published [3], however, their applicability to the domain of cardiac surgery remains unexplored. Cardiac surgery is a highly specialized and complex field of medicine, driven by precision and continuous innovation, which is communicated to peers through surgical lecture videos. In this work, we explore the potential of translating recent advancements in vision-language models to the domain of cardiac surgery. The developed model is applied to the important task of text-based video retrieval, enabling efficient search of relevant content in large unlabeled surgical video databases and potentially enhancing their educational utility.

***Corresponding author: Georgii Kostiuchik, Jimmy Chen, Mina Farag, Gabriele Romano, Matthias Karck, Rawa Arif, Sandy Engelhardt,** Department of Cardiac Surgery, Heidelberg University Hospital, Im Neuenheimer Feld 420, Heidelberg, Germany e-mail: georgii.kostiuchik@med.uni-heidelberg.de
Philippe Grieshaber, Department of Cardiothoracic Surgery, Münster University Hospital, Münster, Germany
***Corresponding author: Georgii Kostiuchik, Jimmy Chen, Mina Farag, Gabriele Romano, Matthias Karck, Rawa Arif, Sandy Engelhardt,** Medical Faculty of Heidelberg University, Heidelberg, Germany

2 Methods

2.1 Data collection and preprocessing

For the training of the vision-language model, a carefully curated set of 229 cardiothoracic video lectures, has been initially collected from two widely-recognized journals and public surgical educational resources in the cardio-thoracic domain – Annals of Cardiothoracic Surgery [4] and The Multimedia Manual of Cardio-Thoracic Surgery [5]. The collected set of videos predominantly comprises heart valve and aorta procedures and covers a broad spectrum of pathologies and surgical techniques.

To leverage the expert surgeon narration of the surgical lecture videos, we transcribe each video using an open-source automatic speech recognition model WhisperX (*medium.en*) [6]. However, due to the complex medical terminology and domain-specific abbreviations, the transcription model is susceptible to misspellings of complex medical terms, omissions, or hallucinations. To enhance the overall quality of audio transcriptions, a set of additional post-processing steps is applied. As such, we eliminate sentences consisting of less than 8 words or span less than 3 seconds, and also excessively long sentences. Furthermore, we manually inspect the transcribed sentences and remove sentences that are not related to a surgical scene, resulting in 8142 narration sentences in total.

There is growing evidence that Large Language Models (LLMs) can be used to enhance textual training data for vision-language tasks [7]. Motivated by the success of these methods, we aim to apply an LLM to enhance textual training data, which currently provides a relatively weak supervision signal due to transcription errors and lack of contextual information.

Open Access. © 2025 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

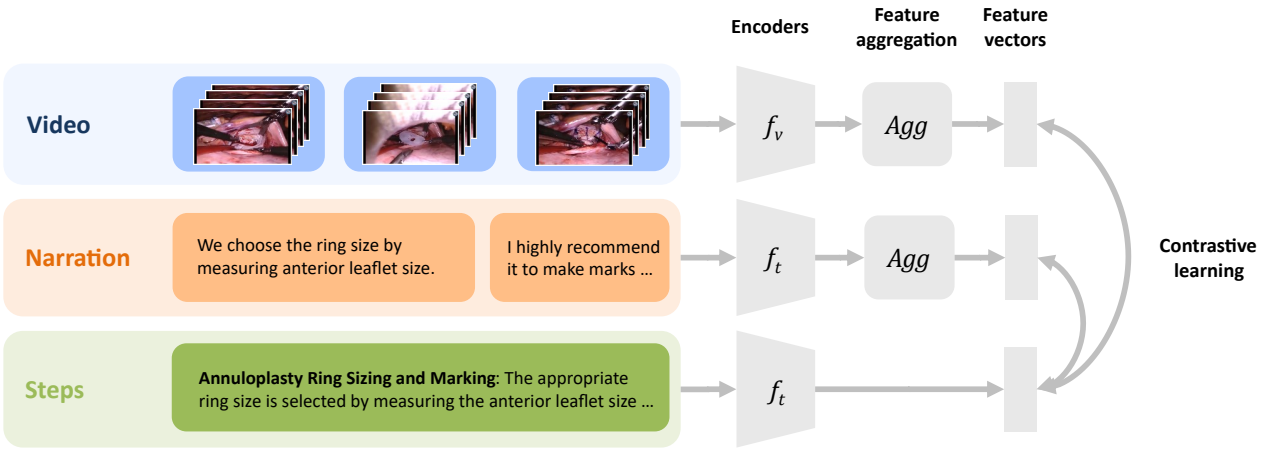


Fig. 1: Overview of the coarse-grained training procedure, adapted from HierVL [11] and PeskaVLP [3]. During coarse-grained training, representations of short-term video clips and narration sentences are averaged and contrasted with the LLM-extracted procedural step. Note that for video clips, we first average frame embeddings within each clip to compute clip-level representations, which are then averaged to produce a coarse video-level representation.

To alleviate these challenges, we propose using GPT-4.1 to extract surgical procedural steps directly from the raw audio transcription text, leveraging the model's strong summarization capabilities and domain knowledge. The motivation for extracting surgical procedural steps from the complete audio transcription of each surgical video lies in the complexity of cardiac surgery procedures, which are characterized by highly patient-specific treatments. Here, we use the following prompt shown in Figure 2 in conjunction with the corresponding audio transcript, resulting in 3345 procedural steps in total. In the next step, we determine the start and end timestamps of the generated procedural steps by aligning them with transcription sentences using a sentence transformer model (*all-mpnet-base-v2*) [8]. More specifically, we calculate the cosine similarity between each procedural step and the transcription sentences, and apply a threshold to identify matching samples, analogous to Li et al. [7].

> *You will receive an audio transcript of a surgical educational video. Your task is to provide a concise summary of each individual step of the surgery, listed in chronological order. For each step, include specific actions performed by the surgeon, a description of the surgical scene (including relevant anatomy, tissues, and organs), as well as the instruments used and techniques applied. Present the information in a manner that an expert surgeon might use to explain their actions to surgical residents during the procedure. The output should consist of the step number followed by the corresponding description. Each step should be clear, self-contained, and understandable on its own. Please use the present tense. Below is the transcript of the surgical video:*

Fig. 2: LLM prompt used for the extraction of procedural steps from raw audio transcripts of cardiac surgery lecture videos.

2.2 Network architecture

The network architecture follows a dual encoder approach, analogous to the CLIP [1] architecture, consisting of an image encoder and a text encoder. To leverage the rich representations of large-scale pretrained models, we adopt EndoViT [9] as image encoder f_v and BioClinicalBERT [10] as the text encoder f_t. Both encoders are extended with a single linear layer which project the vectors into a common embedding space.

2.3 Training objective

During model training, we adopt a hierarchical approach, following HierVL [11] and PeskaVLP [3], which incorporates two training objectives. The first objective focuses on aligning narrations sentences with their corresponding short video clips. Here, we sample individual narration sentences along with their temporally aligned short video clips, consisting of 5 video frames sampled at 1 fps. The sentence–clip pairs are then passed through the respective encoders. Since EndoViT is an image encoder, each individual frame is passed through EndoViT, and the resulting $[CLS]$ embeddings from the final layer are averaged to obtain a representation of the video clip. Subsequently, the resulting $[CLS]$ embeddings from the final layers are passed through separate linear layers to project the vectors into a shared embedding space. We use the symmetric loss formulation from CLIP [1] for fine-grained training.

In addition to sentence–clip training, we adopt a coarse training objective where we sample 3 short-term representations of video clips as well as narration sentences, and are average them to form coarse-grained representations. These

Tab. 1: Text-to-video retrieval results on 535 video-text pairs from the test set. The column "'video'" indicates whether the model has been trained using frames or video-clips. R@K denotes Recall@K metric.

Model	Domain	Video	Training data	R@1 (%) ↑	R@5 (%) ↑	R@10 (%) ↑	Mean rank ↓	Median rank ↓
CLIP [1]	General		-	2.24	4.30	6.36	211.75	192
BiomedCLIP [12]	Medical		-	1.31	5.05	9.53	164.99	135
Our	Surgical		Transcriptions	4.86	10.84	17.38	133.93	92
Our	Surgical	✓	Transcriptions	5.79	11.78	16.45	132.83	84
Our	Surgical	✓	Transcriptions + Steps	6.17	13.27	20.56	110.81	61

coarse representations are then contrasted with the procedural steps. A diagram of the coarse objective is shown in Figure 1. We train the model with video–narration pairs for two consecutive epochs, then switch to the coarse training objective for one epoch, repeating this cycle throughout the training. Similarly, we use the symmetric loss formulation from the CLIP approach [1] for coarse training.

The model is trained for 100 epochs using AdamW optimizer with a learning rate of 1e-4, weight decay of 1e-2, and a warm-up of 1000 steps. The batch size is set to 120 for fine-grained training and 50 for coarse-grained training. A set of image augmentations is applied during training, including rotation, color jitter, and horizontal and vertical flips.

2.4 Experiments

In order to assess the quality of learned representations, we directly apply the trained model to the text-based video retrieval task, addressing the use case of retrieving relevant video clips from large unlabeled surgical video databases. Here, we use a separate set of video lectures as a test set. Since individual sentences may lack sufficient contextual information for retrieval task, we concatenate consecutive narration sentences to form short descriptions spanning up to 20 seconds, resulting in 535 video-narration pairs.

In order to apply the pretrained model to the text-based video retrieval task, we pre-compute embeddings of all video clips and text descriptions in the test set. Next, we calculate the cosine similarity between each video clip and text pair, and rank them accordingly. We report the results of text-based video retrieval in Table 1 using common retrieval metrics, i.e., Recall@1, Recall@5, Recall@10, as well as mean and median rank. The Recall@K metric denotes the percentage of video clips that have been ranked within the top K results (the higher the better). Additionally, we report the mean and median rank of the relevant video clips (the lower the better).

3 Results and discussion

We evaluate the performance of the model across different training datasets. As shown in Table 1, hierarchical training on both transcription sentences and procedural steps yields the most optimal performance, in comparison to the training on the narration data. These results suggest that hierarchical training yields greater supervisory signal. This corresponds with findings from studies in other domains [3, 7].

We attribute this to the fact that the LLM-generated steps provide more contextualized and higher-quality descriptions than raw narration sentences and the hierarchical training allows the model to capture the broader context of narration sentence. Figure 3 presents qualitative examples of text-based video clip retrieval.

In addition, we evaluate the performance of current state-of-the-art models, i.e., CLIP [1] and BiomedCLIP [12], however, these models were not explicitly developed for the surgical domain and therefore should not be directly compared to our model.

The evaluation setup presented here also has certain limitations. Evaluation on diverse and higher-quality datasets are necessary to ensure a more objective assessment of performance. Furthermore, the generalization of the learned representations to surgical vision tasks, such as phase recognition and instrument recognition is deferred to future work.

4 Conclusion

In this work, we presented a preliminary study on the training of a vision-language model on openly available cardiac surgery lecture videos. These preliminary results indicate that augmenting the training data by deriving procedural steps for each surgical case using LLMs can increase the model's performance of the model on text-based video retrieval task. In future work, such models could be developed and applied to intraoperative video data, solving several concurrent surgical data science tasks.

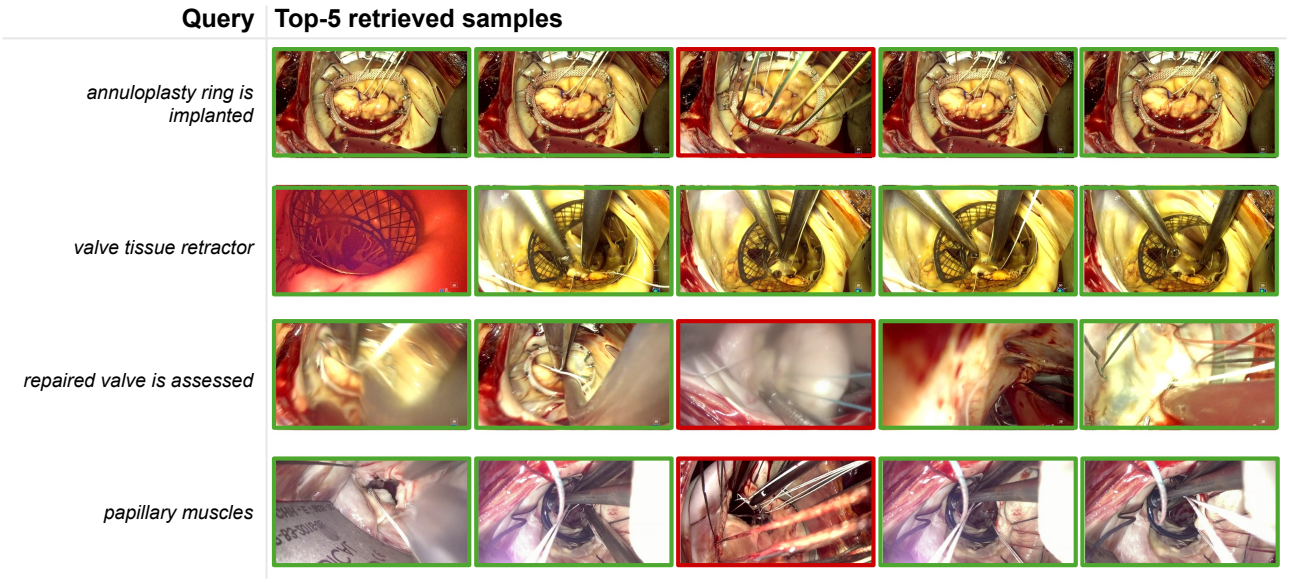


Fig. 3: Qualitative examples of the text-video retrieval capabilities of the model on a private mitral valve repair dataset collected at the University Hospital Heidelberg, comprising 7550 five-second long video clips. The top-5 retrieved samples are highlighted in green or red, indicating whether we consider them correct or incorrect.

Acknowledgment: The authors gratefully acknowledge the data storage service SDS@hd supported by the Ministry of Science, Research and the Arts Baden-Württemberg (MWK) and the German Research Foundation (DFG) through grant INST 35/1503-1 FUGG. The authors also acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG.

Author Statement

Research funding: The authors state no funding involved. Conflict of interest: Authors state no conflict of interest.

References

[1] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning 2021;139:8748–8763.

[2] Yuan K, Srivastav V, Yu T, Lavanchy JL, Marescaux J, Mascagni P, et al. Learning multi-modal representations by watching hundreds of surgical video lectures. Medical Image Analysis. 2025 Oct 1;105:103644.

[3] Yuan K, Srivastav V, Navab N, Padoy N. Procedure-Aware Surgical Video-language Pretraining with Hierarchical Knowledge Augmentation. Advances in Neural Information Processing Systems. 2024 Dec 16;37:122952–83.

[4] Annals of Cardiothoracic Surgery [Internet]. [cited 2025 May 29]. Available from: https://www.annalscts.com

[5] Multimedia Manual of Cardio-Thoracic Surgery [Internet]. [cited 2025 May 29]. The Multimedia Manual of Cardio-Thoracic Surgery. Available from: https://mmcts.org

[6] Bain M, Huh J, Han T, Zisserman A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio [Internet]. arXiv; 2023 [cited 2025 May 29]. Available from: http://arxiv.org/abs/2303.00747

[7] Li Z, Chen Q, Han T, Zhang Y, Wang Y, Xie W. Multi-sentence Grounding for Long-Term Instructional Video. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision – ECCV 2024. Cham: Springer Nature Switzerland; 2025. p. 200–16.

[8] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Internet]. arXiv; 2019 [cited 2025 May 29]. Available from: http://arxiv.org/abs/1908.10084

[9] Batić D, Holm F, Özsoy E, Czempiel T, Navab N. EndoViT: pretraining vision transformers on a large collection of endoscopic images. Int J CARS. 2024 Jun 1;19(6):1085–91.

[10] Alsentzer E, Murphy J, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop [Internet]. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019. p. 72–8. Available from: https://www.aclweb.org/anthology/W19-1909

[11] Ashutosh K, Girdhar R, Torresani L, Grauman K. HierVL: Learning Hierarchical Video-Language Embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023. p. 23066–78.

[12] Zhang S, Xu Y, Usuyama N, Xu H, Bagga J, Tinn R, et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. NEJM AI. 2025;2(1):AIoa2400640.